May 99

RAND

An Assessment of Technical Issues Raised in R.W. Haley's Critique of Three Studies of Health Effects of the Gulf War

Gus W. Haggstrom

National Defense Research Institute

20001027 089

RAND is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND® is a registered trademark. RAND's publications do not necessarily reflect the opinions or policies of its research sponsors.

Preface

The Department of Defense has been conducting a series of studies of health effects of veterans who served in the Persian Gulf War. As part of that effort, RAND has been working with the Office of the Special Assistant to the Deputy Secretary of Defense for Gulf War Illnesses to compile a series of literature reviews and policy papers. Three government-sponsored studies were published in the *New England Journal of Medicine* in 1996 and 1997. These studies were critiqued by R.W. Haley, and this critique was published in *The American Journal of Epidemiology*, along with responses by the authors of the three articles in question and a reply by Haley to those responses. The Special Assistant asked RAND to review R.W. Haley's critique along with the responses to that critique by the authors of the three studies and Haley's reply. This document reports the results of that review. This research was begun in 1998, and a completed draft was provided to the sponsor in May 1999.

This work is sponsored by the Office of the Special Assistant and was carried out jointly by RAND Health's Center for Military Health Policy Research and the Forces and Resources Policy Center of the National Defense Research Institute. The latter is a federally funded research and development center sponsored by the Office of the Secretary of Defense, the Joint Staff, the unified commands, and the defense agencies.

Contents

Summary

In a 1998 article in the *American Journal of Epidemiology*, R.W. Haley challenged the validity of three government-sponsored studies that found that military personnel deployed to the Persian Gulf region in connection with the 1991 Gulf War experienced no excess risk of adverse health effects. The three studies, which were published in the *New England Journal of Medicine* in 1996 and 1997, used multivariate statistical procedures to contrast postwar rates of death, hospitalization, and birth defects among Gulf War veterans with those for other military personnel who were deployed elsewhere. Haley claimed that the authors' statistical methods were flawed and their findings were distorted by various biases. The three authors published rebuttals to Haley and Haley also prepared a response to their reply all in the same issue of the *American Journal of Epidemiology*.

This study undertook a thorough review of the three original studies to examine the technical issues that Haley raised, focusing on his criticisms of the statistical work and the authors' rebuttals. In essence, Haley argues that the studies' authors—in calculating relative health risk ratios for Gulf War veterans and for other veterans—did not account for the fact that the database they used to assess Gulf War illness resulted from a complete sample of Gulf War veterans and approximately a 50% random sample of the nondeployed veterans.

Haley treats the sampling variability in the mortality and hospitalization rates based on these two huge samples as the only source of randomness affecting the relative risk ratio estimates. He maintains that the correct formulas for calculating confidence intervals and for gauging statistical inferences are those tailored to this narrowly specified sampling situation. This review examined the validity of this argument and delineated counterarguments for basing statistical analyses on more general formulations, including the superpopulation models that the studies in question used.

This review concludes that, in the context of assessing adverse health effects based on observational data, Haley's formulation exaggerates the precision of statistical measures, ignores numerous sources of random error affecting these measures, and constitutes an unsatisfactory basis for statistical analyses. Moreover, even if one accepts his calculations, the paper fails to make the case

that revised analyses would invalidate the other studies' overall findings of no adverse health effects linked to Gulf War deployment. While Haley's work alleges that the studies also are distorted by biases that the authors have not properly accounted for, a hard look at this argument reveals little or no basis for this criticism. In sum, this review supports the authors' rebuttals of Haley's criticisms and concludes that they stem mainly from erroneous suppositions and misunderstandings.

AN ASSESSMENT OF TECHNICAL ISSUES RAISED IN R.W. HALEY'S CRITIQUE OF THREE STUDIES OF HEALTH EFFECTS OF THE GULF WAR

Dr. R.W. Haley's critique [1], "Bias from the 'Healthy-Warrior Effect' and Unequal Follow-Up in Three Government Studies of Health Effects of the Gulf War," was published in the August 15, 1998, issue of the *American Journal of Epidemiology*, along with the responses to Haley's paper [2. 3. 4] by the authors of the studies in question [6, 7, 8] and a reply to those responses by Haley [5]. This review attempts to clarify some of the technical issues raised in these commentaries and provides an independent appraisal of the arguments from a statistician's perspective.

The three studies, published in the *New England Journal of Medicine* in 1996 and 1997, examined possible health consequences of Gulf War deployment on military personnel by contrasting postwar rates of death, hospitalization, and birth defects among Gulf War veterans with those for "nondeployed" veterans, i.e., those who served during the same period but did not deploy to the Gulf War. Haley levels several criticisms against the studies:

> A joint review of the three papers . . . indicates that the three studies were strongly biased toward finding no excess risk in the deployed veterans. The biases resulted from errors in the calculation of confidence intervals for tests of statistical significance, a failure to appreciate a more pertinent application of the "healthy-soldier effect," and the unequal effects of excluding hospitalizations in nonmilitary hospitals (p. 315).

As this citation illustrates, Haley makes heavy use of the word "bias" in a colloquial sense. By saying that the studies were "strongly biased toward finding no excess risk," Haley implies that the authors of all three studies skewed their findings to downplay adverse health effects among Gulf War veterans. By asserting biases in analyses, as in the boldface section headings "Bias in analyses of hospitalization" and "Bias in analyses of mortality," he not only impugns the statistical work but also the analysts themselves, alleging that they used faulty methods, incorrect formulas, and erroneous interpretations. Fortunately, Haley backs up his charges with very clear expositions of the statistical issues, and the authors of the three studies have responded in kind, so the allegations can be examined one by one to determine their validity.

CALCULATION ERRORS

Haley's main criticism of the statistical work in the mortality and hospitalization studies [6, 7] was that the authors' analyses and interpretations were flawed because they omitted "finite population correction factors" in calculating confidence intervals for certain "relative risk ratios." The meanings of these statistical concepts and their relevance to Haley's arguments will be spelled out later, but the nub of the issue is whether Haley's formula on p. 316 for evaluating confidence intervals for risk ratios is valid and applicable in this situation. If so, the authors have understated the implied precision of the relative risk ratios in such a way that some statistically significant differences between the hospitalization and mortality rates of Gulf War veterans and those of nondeployed veterans would be interpreted as insignificant. Because Haley cites these technical "errors" to buttress his claims that the authors erroneously discounted the "excess risks" of Gulf War deployment, the validity and applicability of Haley's formula are key issues in understanding the dialogue between Haley and the authors.

The statistical framework that Haley has in mind is a narrowly defined sampling situation in which N_1 persons in a population of N persons have undergone some "treatment" (e.g., Gulf War deployment) and the remaining $N_0 = N - N_1$ have not. Let θ_1 denote the mortality or hospitalization rate over some period among the N_1 treated members, and let θ_0 denote the analogous rate for the untreated group. Then the *relative risk ratio* $\rho = \theta_1/\theta_0$ is the ratio of the two rates. Next, suppose that, to estimate the population parameters θ_1, θ_0, and ρ, one can observe random samples of sizes n_1 and n_0 taken without replacement from the two groups. Then the sample means (proportions) P_1 and P_0 can be used to estimate the population means, and the sample relative risk ratio $R = P_1/P_0$ can be used to estimate the population risk ratio ρ.

As is shown in the appendix, it follows from these assumptions that, for large values of n_1 and n_0, an approximate 95% confidence interval for ρ is given by

$$R \cdot \exp(\pm 1.96 \sqrt{T_1 + T_0})$$

where

$$T_i = \frac{1 - P_i}{n_i P_i} \cdot \frac{N_i - n_i}{N_i - 1}.$$

This formula for the confidence interval endpoints agrees with Haley's formula on p. 316 for the case $\alpha = .05$, except that the denominator $N_i - 1$ in the formula for T_i has been replaced by N_i. The second factor in the formula for T_i is called

the *finite population correction (fpc)* factor. It is noteworthy that the *fpc* factor shrinks to zero (and the confidence interval endpoints merge) as the sample sizes n_i approach the population sizes N_i.

In the present case, where $n_1 = N_1$ and $n_0 \approx N_0/2$ (i.e., all Gulf War veterans were included in the study, and about half of the nondeployed veterans were randomly selected for inclusion), the term T_1 vanishes and T_0 is only about half as large as it would be if the *fpc* factors were omitted, so that the inclusion of the *fpc* factors leads to markedly shorter confidence intervals. An implication of this observation for statistical inferences follows from the fact that, given a 95% confidence interval for a relative risk factor, one can test the hypothesis of no differences in the population proportions θ_i (or equivalently that $\rho = 1$) at the 5% significance level by observing whether or not the confidence interval covers the value 1.0. Thus, the inclusion of the *fpc* factors in calculating several confidence intervals, say, for rates of mortality attributable to various causes of death, could lead to identifying more statistically significant differences than if the *fpc* factors were omitted.

To sum up, Haley's formula for a 95% confidence interval for θ_1/θ_0 is valid in a narrowly defined sampling situation. However, this confidence interval is for a ratio of population *means*, θ_1/θ_0, not for "adjusted" means. Unless the treatment and untreated groups are balanced on the key risk factors related to the outcomes of interest (to assure that θ_1 and θ_0 would be about the same if it were not for the effects directly attributable to Gulf War deployment), there is no basis for inferring that sample risk ratios P_1/P_0 considerably greater than (or less than) 1.0 reflect the magnitudes of the treatment effects.

THE "HEALTHY WARRIOR" EFFECT

In the three studies that Haley criticizes, the crude sample risk ratios P_1/P_0 were presented in conjunction with detailed multivariate analyses to control for differences between the Gulf War veterans and their nondeployed counterparts in risk factors associated with the health-related outcomes under investigation. In this case, there were marked differences between the two groups in terms of personal attributes (age, race, gender, type of unit, occupational specialty)—and presumably in unmeasured health status measures as well—that had to be accounted for in gauging the effects of Gulf War deployment. Haley clearly understands the need to adjust the crude relative risk ratios for these differences, arguing that the treatment and untreated groups are definitely not comparable here because of what he calls the "healthy-warrior effect."

Statisticians have a wide variety of models and methodologies (e.g., loglinear models, logistic regression, proportional hazards models, analysis of covariance, indirect or direct standardization) for incorporating "covariates" and risk factors into analyses of this type. These methodologies provide "adjusted" rates for both groups and adjusted risk ratios analogous to P_1/P_0 that control for differences in risk factors between the groups. If the covariates can be shown to be irrelevant to the assessment of the treatment effects, then the adjusted risks would coincide with the unadjusted risks, and the adjusted relative risk ratios would be the same as the crude ratios P_1/P_0. In reporting the results of such an analysis, most applied statisticians would report confidence intervals and standard errors for the crude ratios that are consistent with their modeling assumptions and analytic framework. And, since most statisticians use *superpopulation models* in which the individual observations are treated as being independent (or at least uncorrelated), the *correct* formula for the confidence intervals endpoints, under those assumptions, would *omit* the finite population correction factors. This is the essence of the authors' responses to Haley on this matter.

Superpopulation Models

Thus, the applicability of Haley's formula depends in part on the tenability of superpopulation models and "model-based" analyses. Haley argues that superpopulation models are inappropriate here, because "Gulf War veterans constitute a unique, finite population, one that has never existed before, for which the defining circumstances are unlikely to recur, and for which we can identify all members." Gray et al. argue otherwise, conceding that the choice between finite population and superpopulation models is a matter of debate among some statisticians:

> Briefly, this philosophical debate as applied to this study concerns whether one wishes to consider the hospitalization experience of Gulf War veterans and nondeployed veterans to be one deterministic experience (finite population model) for which we have complete data or one realization of a stochastic experience (superpopulation model). . . Under the finite population model, there is essentially no random variability, except that the nondeployed veteran population was sampled at a 50 percent rate, which results in (essentially) null confidence intervals. Under the superpopulation model, there is stochastic variability, and the confidence intervals reported in the hospitalization paper apply (p. 328).

Actually, there is little debate among applied statisticians on this issue. They routinely adopt superpopulation models, which are commonly referred to as "survival models" or "hazard rate models" in this context, as a basis for

analyzing categorical data that arise from counting processes of the types considered here, namely, counts of deaths, accidents, illnesses, hospitalizations, birth defects, etc., during some time interval (see references 5 through 8). In applications of these types, the models commonly adopted reflect variability in the severity, timing, classification, and resolution of the health-related episodes underlying the cell counts, and the counts themselves are treated as realizations of stochastic processes (e.g., time-dependent Poisson processes). Does the adoption of these models affect the calculation of standard errors and confidence intervals for relative risk ratios? The answer is that, if the population means θ_i are small (as they are here), the net effect of assuming that the hospitalization or mortality category counts have Poisson distributions (in lieu of hypergeometric distributions) is to *omit* the *fpc* factors in the formula for T_i.

To support his contention that *fpc* factors are required, Haley cites W.G. Cochran's monograph on sampling techniques [13], which is restricted primarily to simple sampling situations. Nevertheless, Cochran makes it clear that he views superpopulation models as viable alternative frameworks for analyzing complex survey data; in Sections 6.7 and 7.8, he shows the concurrence of "design-based" (finite population) estimators with best linear unbiased estimators in simple linear regression models, and he notes the simplicity and exactness of model-based variance calculations. For more general discussions pointing to the tradeoffs between model-based and design-based methodologies for complex sampling designs, see references 14 and 15.

Analysis of Mortality Rates

Haley presents his calculations of confidence intervals for relative mortality rate ratios in Table 2, a key table in the dialogue because it constitutes the basis for Haley's claim that the postwar mortality rates are distorted by "selection biases" due to the healthy-warrior effect. To separate issues here, I have checked that the 95% confidence intervals for the mortality rate ratios listed in the table are consistent with the reported numbers of deaths and sample sizes, so the correctness of Haley's confidence intervals is not at issue here. However, Haley barely mentions these confidence intervals in his discussions of the rates. Perhaps by listing them under pairs of crude and adjusted rates, he may have intended to invite the reader to conjecture that the adjusted rates would have about the same relative precision as the crude rates, but there is no basis for that conjecture.

To make his case, he argues that the *pattern* of the crude, cause-specific mortality rate ratios in Table 2, in conjunction with the adjusted mortality rate ratios taken from the Kang and Bullman study [6], indicates excess postwar mortality among

Gulf War veterans. He reasons that the very low crude mortality rate ratios for infectious diseases (0.22), cancers (0.59), diseases of the digestive system (0.61), and diseases of the circulatory system (0.87) show that the "magnitudes of the selection biases were large." He goes on to assert on p. 319 that, given the pattern of the crude and adjusted cause-specific mortality rates, other ratios that are near 1.0 or larger would be substantially higher if it were not for the healthy-warrior effects that are not fully accounted for in the Kang-Bullman analysis:

> The rate ratios were close to 1.0 for death from diseases of the respiratory system, suicide, and homicide and substantially greater than 1.0 for death from motor vehicle accidents (Table 2). Since the "healthy-warrior effect" must have included an excess of personnel with chronic respiratory illness and major depression in the nondeployed group, postwar mortality rate ratios near 1.0 suggest that the deployed group suffered excess postwar death from respiratory illness. . . . In contrast, since death from homicide and motor vehicle accidents is not known to have antecedents that would prevent a soldier from being deployed to the war zone (no "healthy-warrior effect"), *their postwar rate ratios are probably unbiased estimators of the true excess mortality risk due to deployment.* (Italics added.)

This argument has several holes. First, Haley exaggerates the potential selection biases from the healthy-warrior effect. Challenging Haley on this score, Kang and Bullman contend that "the effects of the potential selection bias on the mortality outcomes are minimal and negligible" (p. 325), and they present a table showing a remarkable concordance between the cause-specific mortality rates of activated reservists and those for unactivated reservists during the 1991–1993 postwar period, thereby refuting Haley's contention. Gray et al. also challenge the basis for Haley's supposition and present additional information, notably Figure 1, to show that the selection effect on hospitalizations was "transient and largely resolved by the conclusion of the conflict" (p. 328).

Second, Haley glosses over the fact that he is dealing with very small numbers of deaths relative to the huge sample sizes. Because the cause-specific death rates are tiny, the ratios are suspect not only because of their tiny denominators but also because of possible classification errors in the cause-specific death counts. In particular, Haley's comment in the paragraph cited above regarding deaths from diseases of the respiratory system was based on just 14 deaths among the 695,516 deployed veterans and the same number among the 746,291 other veterans in the sample. Thus, the crude mortality rates for the two groups are 0.0000201 and 0.0000188, and the relative risk ratio is 1.07, for which Haley reports a 95% confidence interval of (0.74, 1.56), as compared with my calculation of (0.51, 2.25) when the *fpc* factors are omitted. Kang and Bullman in [6, p. 1499] reported the

analogous 95% confidence interval for the adjusted mortality rate ratio, 1.27, to be (0.60, 2.70). No matter which interval estimates are used, it is hard to see how Haley could infer that these numbers "suggest that the deployed group suffered excess postwar death from respiratory illness."

While the death counts for other disease-related causes are somewhat higher, the cell counts are still very small, and even the confidence intervals that Haley provides are wide enough to raise doubts as to whether Table 2 provides evidence supporting Haley's argument. Moreover, the cell counts themselves may not be reliable, given that the causes of death were determined from death certificates that may have misreported the principal cause of death or may have reported multiple causes. Kang and Bullman concede possible classification errors in their study, asserting that "death certificates dependably establish the fact of a person's death, but their accuracy in recording the cause is variable" [6, p. 1503]. Given that the causes of death were then computerized using ICD-9 codes (from the *International Classification of Diseases, 9th Revision*) and given the possibilities of coding and recording errors, not only in the ICD-9 codes, but also in the Social Security numbers of the deceased veterans and the unit designation codes used to classify the veterans into the deployed and nondeployed groups, there is room to question the reliability of the cell counts.

If there is fuzziness in the categorization of the causes of the disease-related deaths, and if there is reason to expect that the effects of Gulf War deployment on mortality rates might be relatively uniform over categories, then those effects would manifest themselves in the mortality rates and ratios derived from the pooled counts over all disease-related categories. There were 337 deaths from all disease-related causes among the 695,516 Gulf War veterans and 534 among the 746,291 nondeployed veterans, so that the crude mortality rate ratio was 0.68. While that would seem to support Haley's case for sizable healthy-warrior effects, Kang and Bullman report that the analogous adjusted mortality rate ratio for the pooled counts was 0.88, and the associated 95% interval estimate was (0.77, 1.02), which provides weak evidence to support Haley's supposition of healthy-warrior effects and no evidence of excess deaths attributable to Gulf War deployment.

Deaths Attributable to External Causes

The death counts attributed to external causes are larger, but there were only 1,765 deaths from all causes among Gulf War veterans and only 1,729 in the comparison group, so that the overall mortality rates were 0.00254 and 0.00232. A substantial majority of those deaths, 1,317 and 1,081, were attributed to

accidents, suicides, and homicides, so their linkages to Gulf War deployment are questionable.

Of the three externally caused death categories that Haley lists in Table 2, deaths attributed to motor vehicle accidents were the most numerous—549 Gulf War veterans versus 398 other veterans—and yielded the highest mortality rate ratio, 1.48, for which the associated 95% interval estimate is (1.38, 1.59) if the fpc factors are included, and (1.30, 1.69) otherwise. Since the analogous adjusted mortality rate ratio and interval estimate reported in [6] are 1.31 and (1.14, 1.49), I see no basis for Haley's claim that the 1.48 figure represents a "probably unbiased" estimate of the "true excess mortality risk due to deployment." In any case, the crude and adjusted mortality rate ratios for deaths due to accidents are quite high, and they raise further questions as to whether other factors must be considered in assessing the effects of Gulf War deployment.

What was there about Gulf War deployment that might account for about 150 more deaths from motor vehicle accidents during the three-year period 1991–1993? And how can one explain the high relative risk ratios for other accidents and the significantly higher externally caused death rates for *female* Gulf War veterans [6, p. 1501]? Although Kang and Bullman dismiss their findings of higher risk ratios for males as being statistically insignificant based on their Cox's regression analyses and they cite previous studies finding increased postwar mortality from accidents for veterans from previous wars, their explanation seems unsatisfactory.

Alternative Explanations

A more plausible explanation for these findings is that they reflect what might be termed "separation effects." According to Table 1 in the hospitalization study [7], the Gulf War veterans separated from service at a considerably higher rate (42.5%) through 1993 than those in the comparison group (36.4%). Applying those rates to the numbers of Gulf War veterans and other veterans in the mortality study (695,516 and 746,291), I infer that approximately 296,000 of the Gulf War veterans had returned to civilian life through 1993, outnumbering the corresponding figure (241,000) for other veterans by 22%.

One implication of the difference in separation rates is that more Gulf War veterans underwent separation physical examinations. Haley alleges that some veterans would fail to report serious illnesses on those exams, but Gray et al. defend the rigor of the exams and argue that the veterans had considerable incentives to report their medical problems fully (p. 330). If these exams led to identifying and treating some of the serious medical conditions, thereby preventing later complications and deaths, this would supplant healthy-warrior

effects as an explanation for the somewhat lower disease-related mortality rates for Gulf War veterans than for other veterans. The adjusted mortality rate ratio for disease-related deaths was 0.88, and the 95% interval estimate was (0.77, 1.02).

Another implication of the difference in separation rates is that many more Gulf War veterans became subject to the perils of civilian life. If we assume for the moment that the separatees, being free of military constraints on personal behavior and living under less-controlled environments, would have higher mortality rates due to external causes (accidents, suicide, and homicide), then one would expect proportionately more deaths in these categories among Gulf War veterans than among other veterans. The premise that civilian life is more hazardous is partially substantiated by Table 4 in [6], which shows that the standardized mortality ratios for all external causes were 0.64 for the Gulf War veterans and 0.55 for other veterans. Thus, even though the death counts for Gulf War veterans through 1993 included large numbers of deaths after they had returned to civilian life, and they included the relatively high death counts for the women in this group, the Gulf War veterans still had 36% fewer deaths than one would predict based on mortality rates for civilians having the same age, sex, and race attributes.

Based on the standardized cause-specific mortality rate ratios in Table 4 of [6] and Kang and Bullman's citation of a study of U.S. Army soldiers showing that the mortality rate of soldiers in 1986 was only half the rate of their civilian counterparts (p. 1503), I conjecture that a reexamination of the mortality data would support the hypothesis that there was a sizable jump in the hazard rate (force of mortality) for externally caused deaths at the separation date. If analyses of the timing of deaths among both groups of veterans support the conclusion that the hazard rates for external causes were about twice as high after the separation date than they were before, that finding, in conjunction with the higher separation rates among Gulf War veterans, would account for the higher adjusted mortality rate for all external causes among male Gulf War veterans, and it might even account for the very high mortality rates for the women who served in the Gulf.

Carrying this argument another step, suppose that it can be shown that both groups of veterans had exactly the same cause-specific mortality rates while they remained on active duty and they had the same (greatly elevated) mortality rates after they separated. In addition, suppose that it can be shown that Gulf War deployment caused higher separation rates among Gulf War veterans. Then it would follow that the higher mortality rates among Gulf War veterans would be attributable to Gulf War deployment, even though Gulf War deployment had no

effect whatsoever on mortality rates on veterans either before and after they left the service. The point of this argument is not to explain away the differences in the mortality rates, but to pinpoint another salient factor, in addition to Haley's healthy-warrior effect, that merits consideration in weighing the mortality rate ratios.

Of course, not all veterans experienced equally hazardous conditions before or after separation, and it seems reasonable to expect that, if there is, *on average*, a doubling of the hazard rates at the separation point, then there might be no increase whatsoever for many veterans but a tenfold or hundredfold increase for others, say, a female military policeman who leaves the service to become a inner-city cop. When one considers the extent to which individuals vary in their lifestyles, environments, and exposures to hazards, a case can be made for more detailed analyses of the mortality data to account for variability in those factors, but I see nothing in the Kang-Bullman study indicating that they glossed over important risk factors in their assessment. In fact, I see numerous details in their report indicating that they strove to turn up adverse effects of Gulf War deployment. Witness their telling statement on p. 1502 in [6]: "Of the 10 deaths attributed to infectious or parasitic disease, none were reported as due to leishmaniasis or other infectious diseases endemic to the Middle East, or as due to the effects of biologic warfare agents."

THE UNEQUAL FOLLOW-UP ISSUE

Haley's primary criticism of the hospitalization and birth defects studies was that they were distorted by "biases from unequal follow-up." He makes his case as follows:

> Whereas virtually all deaths were equally ascertained in both comparative populations for the first study, the records of hospitalizations, births, and birth defects for the second and third studies were obtained only from military hospitals serving personnel remaining on active duty; the hospital records of personnel who separated from active duty during the follow-up period and were treated in nonmilitary hospitals were excluded (p. 315).

This is a valid criticism, especially in the case of the birth defects study [8], which would have to be a long-term study to be conclusive, perhaps requiring follow-ups of the two veteran populations for ten or twenty years. However, as the preceding discussion of separation effects shows, analysts would have to separate the effects of Gulf War deployment from those of other salient health-related factors. Given the complexity of that task and, perhaps more important,

given "the absence of a clearly defined hypothesis regarding measurable exposures and specific birth effects" (p. 327), I share the authors' concerns about the practicality of undertaking a long-term study. In my view, Cowan et al.'s analyses in [8] and their responses to Haley's criticisms are well-conceived and well-documented.

Insofar as the hospitalization study [7] is concerned, it is not clear that the restriction to military hospitalizations is a serious shortcoming. In fact, there would seem to be some advantages from an analytic viewpoint, since this restriction assures greater uniformity in the reporting of hospitalizations. While Gray et al. have taken steps to augment their database with computerized hospitalization records from the state of California and the Department of Veterans Affairs (p. 330), it will be difficult, if not impossible, to gauge the effects of Gulf War deployment on postservice hospitalizations allowing for the multitude of separation effects that one can hypothesize. Moreover, it seems unlikely that additional postservice data will reveal significant effects of Gulf War deployment on the health outcome measures under consideration, given that those effects have not manifested themselves in either the restricted hospitalization data or the unrestricted mortality data for both groups of veterans through September 1993. Of course, one cannot rule out the possibility that additional follow-up data might facilitate identifying less serious effects of Gulf War deployment in the form of illnesses that do not rise to the level of requiring hospitalization, but those illnesses are not the subject of the studies under review.

Among the specific causes of hospitalizations for which the relative risk ratios were high, the high counts of hospitalizations for mental disorders stand out. If one assumes that the criteria for distinguishing, say, personality disorders from neurotic disorders are somewhat fuzzy, then one must also concede that some higher risk ratios might stem from classification errors that distort the subcategory counts, especially when the counts are as small as they are in these studies. Perhaps the significantly higher risk ratio for hospitalizations in 1992 due to *adjustment reactions* might be attributable to Gulf War deployment, but the vagueness of the categorization raises more questions about the reliability of the counts. (According to the ICD-9-CM manual, adjustment reactions "are usually closely related in time and content to stresses such as bereavement, migration, or other experiences.") In any case, there is only weak evidence of excess hospitalizations for Gulf War veterans in the reported standardized risk ratios. To account for higher relative risk ratios for genitourinary disorders, the authors explain that "the observed differences between cohorts with regard to rates of

diagnoses suggest that medical care for some conditions was deferred until after the war" [7, p. 1511].

Gray et al. provide a multifaceted response to Haley's criticism that the hospitalization study did not fully account for the healthy-warrior effects. First, they challenge the basis of Haley's supposition that there were marked differences between the deployed and nondeployed groups in the prevalence of chronic diseases. Second, they point to their efforts to control prewar selection effects in their analysis by including a "surrogate health status covariate, prewar hospitalization, as a statistical adjustment" (p. 328). Third, they report the results of additional analyses that address the specific issue of healthy-warrior effects.

To carry out the additional analyses, they exploited a virtue of the hospitalization database in that it covered a nearly five-year period from November 1988 through September 1993, thereby permitting comparisons of hospitalization rates between the deployed and nondeployed groups before, during, and after the Gulf War. To allow for the fact that the data were right-censored for veterans who separated from service before September 1993, they employed Cox's regression procedure to estimate the effects of Gulf War deployment on first hospitalizations. Then they extended their analysis to include second and later hospitalizations by using logistic regression to compare the probabilities of hospitalization for deployed and nondeployed personnel during each three-month interval from November 1988 to September 1993. They found that Gulf War veterans experienced slightly lower hospitalization rates prior to deployment, lower rates during the military build-up and conflict (from August 1990 to July 1991), and slightly lower rates after July 1991. Averaging the estimated quarterly hospitalization rates across time intervals, they found that the average probability of hospitalization for Gulf War veterans before August 1990 was 0.0194, while after August 1990 it was 0.0189, whereas the comparable averages for the nondeployed veterans were 0.0218 and 0.0235. They concluded that there was a selection effect stemming from the fact that "only the most fit service members were deployed; however, this effect was transient" (p. 329).

To a certain extent, these findings substantiate Haley's position that healthy-warrior effects exist and must be accounted for in definitive analyses of Gulf War illness. However, the evidence from the three studies indicates that Haley has exaggerated the importance of healthy-warrior effects, and he has downplayed the efforts that the authors undertook to account for those effects. Also, while Haley's criticisms pertaining to unequal follow-up in the hospitalization and birth defect studies are valid, the authors of those studies have responded fully to his criticisms, citing additional analyses to support their findings and interpretations. Of course, questions remain as to whether Gulf War veterans

have experienced significantly more chronic illnesses that do not ordinarily entail hospitalization and whether they have suffered or will suffer more negative long-term health outcomes than their nondeployed counterparts. The clear message that emerges from this review is that Haley's concerns about healthy-warrior effects and unequal follow-ups must be addressed in studies that attempt to answer those questions.

Appendix
DERIVATION OF CONFIDENCE INTERVALS FOR RISK RATIOS

Let $\rho = \theta_1/\theta_0$ denote the risk ratio of interest, where θ_1 and θ_0 are the population proportions to be estimated by sample proportions P_1 and P_0 based on random samples of sizes n_1 and n_0 taken without replacement from populations of sizes N_1 and N_0. Then we know from sampling theory results [13, p. 51] that P_i is unbiased for θ_i with variance

$$Var(P_i) = \frac{\theta_i(1 - \theta_i)}{n_i} \cdot \frac{N_i - n_i}{N_i - 1}$$

and that P_i is asymptotically normal for large values of n_i and N_i. It follows that $Z = \ln R = \ln(P_1 / P_0) = \ln P_1 - \ln P_0$ is asymptotically normal with mean $\ln \rho$ and variance $\sigma_Z^2 = Var(\ln P_1) + Var(\ln P_0)$. Hence, if σ_Z were known, the end points $Z \pm 1.96\sigma_Z$ would define a 95% confidence interval for $\ln \rho$ so that a 95% confidence interval for ρ would be given by

$$\exp(Z \pm 1.96\sigma_Z) = (P_1 / P_0)\exp(\pm 1.96\sigma_Z) = R \cdot \exp(\pm 1.96\sigma_Z)$$.

Moreover, the same asymptotic result holds if one replaces σ_Z by the standard error $\hat{\sigma}_Z$ derived by replacing population means by sample means in the formula for σ_Z^2. Applying the Taylor's formula linear approximation for $f(x) = \ln x$ around $x = \theta$, namely,

$$f(x) \approx f(\theta) + f'(\theta)(x - \theta) = f(\theta) + (1 / \theta)(x - \theta)$$,

we first use the representation $\ln P_i \approx \ln(\theta_i) + (P_i - \theta_i) / \theta_i$ to approximate

$$Var(\ln P_i) = (1 / \theta_i^2)Var(P_i) = \frac{1 - \theta_i}{n_i\theta_i} \cdot \frac{N_i - n_i}{N_i - 1}$$,

which leads to $\hat{\sigma}_Z = \sqrt{Est.\,Var(Z)} = \sqrt{T_1 + T_0}$ where T_i is specified by the formula on page 2.

References

1.   Haley RW. Bias from the "Healthy-Warrior Effect" and Unequal Follow-Up in Three Government Studies of Health Effects of the Gulf War. *Am J Epidemiol 1998*; 148: 315–323.

2.   Kang HK, Bullman TA. Counterpoint: Negligible "Healthy-Warrior Effect" on Gulf War Veterans' Mortality. *Am J Epidemiol 1998*; 148: 324–325.

3.   Cowan DN, Gray GC, DeFraites RF. Counterpoint: Responding to Inadequate Critique of Birth Defects Paper. *Am J Epidemiol 1998*; 148: 326–327.

4.   Gray GE, Knoke JD, et al. Counterpoint: Responding to Suppositions and Misunderstandings. *Am J Epidemiol 1998*; 148: 328–333.

5.   Haley RW. Countercounterpoint: Haley Replies. *Am J Epidemiol 1998*; 148: 334–338.

6.   Kang HK, Bullman TA. Mortality Among U.S. Veterans of the Persian Gulf War. *N Engl J Med 1996*; 335: 1498–1504.

7.   Gray GC, Coate BD, Anderson CM, et al. The Postwar Hospitalization Experience of U.S. Veterans of the Persian Gulf War. *N Engl J Med 1996*; 335: 1505–1513.

8.   Cowan DN, DeFraites RF, Gray GC, et al. The Risk of Birth Defects Among Children of Persian Gulf War Veterans. *N Engl J Med 1997*; 336: 1650–1656.

9.   Agresti A. *Categorical Data Analysis*. New York: Wiley, 1990.

10. Cox DR and Oakes D. *Analysis of Survival Data*. New York: Chapman and Hall, 1984.

11. Elandt-Johnson RC and Johnson NL. *Survival Models and Data Analysis*. New York: Wiley, 1980.

12. Chiang CL. *Introduction to Stochastic Processes in Biostatistics*. New York: Wiley, 1968.

13. Cochran WG. *Sampling Techniques*. 3d ed. New York: Wiley, 1977.

14. Sarndal CE. On π-Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling. *Biometrika 1980*; 67: 639–650.

15. Smith TMF. Present Position and Potential Developments: Some Personal Views—Sample Surveys. *J Royal Statist Soc A 1984*; 147: 208–221.